基于游戏的心理测评：概念，范式及实践

徐俊怡　李中权

（南京大学社会学院心理学系，南京 210023）

摘要 基于游戏的心理测评是指通过游戏或游戏化的活动，对一个人的能力、人格等心理特性和行为进行量化测评。早期主要以评估教育、训练效果为目的而后发展成对心理特性的测评，基于游戏的测评作为一项新技术在测评形式、测评过程和测评结果上均具有优势。目前基于游戏的测评形成了以证据中心设计为基础的范式，用于指导建立测评工具并开展实证研究，在测评个体认知能力和非认知能力方面均有实践。然而当前该技术仍处于起步阶段，未来研究可以在任务设计、结果分析及实践应用方面进一步拓展深入。

关键词 基于游戏的测评，证据中心设计，认知能力，非认知能力

1 引言

根据中国互联网络信息中心发布的《中国互联网络发展状况统计报告(2019)》显示，截止 2019 年 6 月，中国网络游戏网民规模达到 4.94 亿，游戏用户主要集中在 10 至 39 岁，游戏已经成为人类社会行为的重要组成部分。以往研究者将研究重点放在游戏对个体心理及行为的影响上，然而随着大数据时代的到来，由数据带来的革命在各个领域悄然兴起，也为心理学研究的开展提供了新思路。相比于传统的行为测量方式，在大数据时代人们的行为一定程度上可以通过数据来衡量(Schoedel et al., 2018)，因此近年来，如何使用游戏得到丰富的数据并预测玩家的知识、技能和特质受到越来越多关注。

商业领域已初步尝试将游戏元素与心理测验结合起来应用于企业招聘(杨振

收稿日期：2020-05-24

* 教育部人文社科规划基金项目(20YJA190004)；江苏省教育厅高校哲学社会科学研究重点项目(2018SJZDI203)。

通信作者：李中权, E-mail: zqli@nju.edu.cn

芳,孙贻文, 2015)，并开发了一系列游戏系统。Arcitc Shore 公司作为游戏化招聘的先行者，率先使用行为任务来判断应聘者的人格特质；普华永道与其联合建立了 Career Unlocked 的游戏化招聘系统并已投入使用，其中涉及充气球存钱、情绪判断等多项游戏任务；德勤也打造了自己的测评手游 Firely Freedom，通过多个游戏关卡对应聘者的勤奋、完美主义倾向、风险管理与规避能力等特质进行评价。

基于游戏的心理测评拓展了心理测量的手段，与机器学习方法的进一步结合使其在大数据时代拥有巨大的应用潜力，但是作为一个新兴交叉研究领域，基于游戏的测评处于“实践先行，理论滞后”的阶段。目前基于游戏的测评技术未得到国内学者们的广泛关注，因此本

收稿日期：2020-05-24
* 教育部人文社科规划基金项目(20YJA190004)；江苏省教育厅高校哲学社会科学研究重点项目(2018SJZDI203)。
通信作者：李中权, E-mail: zqli@nju.edu.cn

文主要在国外已有研究的基础上，结合少量的国内研究，对基于游戏的测评这一项新技术进行介绍，综述其概念、评估范式和实践应用，并提出未来研究方向，希望为后续研究提供参考。

2 概念评述

2.1 概念发展与界定

游戏本身的含义，是指人们参与交互的一种娱乐方式，娱乐是其具有的本质特征(吴宇, 2015)。21 世纪以来，游戏的巨大潜力受到学者们的关注，它的使用目的不再局限于娱乐，严肃游戏(Serious Game)的概念在国内外兴起。严肃游戏是指通过游戏的娱乐形式达到教育、训练和治疗等严肃的目的(Gamberini et al., 2009)，基于游戏的学习作为严肃游戏的一个分支目的在于让人们从游戏中习得知识和行为(Gee, 2008)。早期基于游戏的测评大多以评估特定的学习结果和技能为出发点，在游戏环境中开发测评模型(Mislevy et al., 2012)，随着基于游戏的测评深入发展，研究者开始将游戏与能力特征联系起来，让游戏提供个体如何思考和行动的线索。

Heinzen 等人(2015)将基于游戏的测评(Game-based assessment, GBA)定义为通过游戏或游戏化的活动，对某一对象进行评估。从心理学视角来说，这就是指采用游戏的方式，对一个人的能力、人格等心理特性和行为进行量化测评(孙鑫, 黎坚, 符植煜, 2018)。根据测评的形式可以分为外部测评(External assessment)和内部测评(Internal assessment)两类，一方面测评可以基于游戏之外的证据，例如，个体最终的解决方案，在口头陈述或自我报告中阐述的理由等(Caballero-Hernández et al., 2017)；另一方面测评可以作为游戏的一部分嵌入游戏中，也称之为隐形评估(Stealth assessment)，即在游戏中嵌入有效的测验衡量个体在游戏环境中的表现(Shute, 2011)。

基于游戏的测评与游戏化测评(Gamification in assessment)的概念十分相似，

都是一种将游戏机制应用于非游戏环境的方式(Attali & Arieli-Attali, 2015)，但两者最重要的区别在于引入游戏机制的目的。游戏化测评发挥作用的前提是通过游戏产生积极的内部激励作用，使个体与特定环境发生长期互动，提高个体参与度和接受度(Nicholson, 2015)，且对个体的测评表现产生积极作用，因此游戏化测评多用于教育领域，目的在于创造一个有利的环境。而基于游戏的测评目的在于根据受测者在游戏中的行为表现数据，对个体的一个或多个特质进行测量与评价，重点在于实现评估的目的。

2.2 优缺点评述

近年来，基于游戏的测评之所以能成为一种比较受欢迎的评估方式，是因为相比于传统的心理测评，其在测评形式、测评过程和测评结果上都具有一定优势。

第一，从测评形式来说，创设了一个真实度比较高的环境，可以通过复杂的任务测量个体对知识和技能的应用情况(Shute et al., 2016)，不同于传统能力测试采用再认、回忆信息或自我报告等方法，基于游戏的测评可以通过设置场景给学生展现其理解和应用知识的机会。此外，基于游戏的测评还可以设置多个关卡，考察个体在不同类别或不同难度的情境下的表现，形式更加灵活。

第二，从测评过程来说，基于游戏的测评可以降低测评过程中的焦虑，提高参与度，得到受测者更为真实的情况。相比于传统心理测评存在测验焦虑、社会赞许性等问题，已有多项研究表明，受测者认为基于游戏的测评方式更有吸引力，趣味性更强(DeRosier & Thomas, 2019; Turan & Meral, 2018)。此外，有研究者将所要考察的题目嵌入游戏对学生进行测试，发现在基于游戏的测评中，学生的考试焦虑有所降低且考试成绩明显更好(Mavridis & Tsiatsos, 2017)。而且基于游戏的测评具有隐蔽的特点，受测者无法猜测测验意图，可以有效减少测

验作假。

第三，从测评结果来说，基于游戏的测评是一种动态连续的过程，可以通过计算机过程数据追踪技术得到受测者在游戏过程中的表现情况，而传统心理测验只能得到最终的结果分数。通过与机器学习例如贝叶斯网络方法相结合可以进一步建立动态变化的模型，并根据受测者表现情况更新测试结果，得到更加准确的数据(Shute et al., 2016)。

但是，基于游戏的测评也不可避免存在一些缺点。对研究者和使用者来说，在测评游戏的开发，测评数据的分析和测评结果的效度三方面均有不少挑战。

从测评游戏的开发来说，通常需要将游戏机制、游戏内容和内容评估结合在一起，由研究人员、游戏设计师和教育工作者等多方参与，共同制作一款专门的游戏。这种方式投入的时间、金钱和人力成本比较高。早期研究多基于现有的一些商业化游戏开展，比如植物大战僵尸等，然而这些游戏本身并不是为评估某种心理特质而开发的，只能针对比较有限的主题进行评估，内容不准确和不完整。也有研究者尝试设计通用性的游戏框架，将内容与游戏机制分离，开发特定主题的游戏，降低游戏开发的门槛，但仍存在一些其他问题，比如心流体验中断、练习效应等(Baron, 2017)。

从测评数据的分析来说，通过基于游戏的测评将收集到大量过程数据，比如鼠标点击次数、反应时间等，一方面这些数据的记录、处理与分析远比传统心理测验得到的数据复杂，这对研究者的数据分析能力提出了较高要求；另一方面，基于游戏的测评关键在于建立过程数据与所测特质结构的关系，如何在众多数据中确立并验证数据指标与所测特质的因果关系对研究者来说具有较大困难(Kim & Ifenthaler, 2019)。

从测评结果的效度来说，基于游戏的测评也存在与传统心理测验一样的问题。有研究者指出，基于游戏的测评的结果并不能完全等于受测者所测特质的实际水平，即使游戏中的任务反映了所测特质的关键要素，受测者如何在游戏中扮演自己的角色并做出一系列行为，只是他们在实际生活中的近似表现(StĂNescu et al., 2020)。而且游戏过程中界面的颜色、角色的造型、游戏的音效等环境要素，以及受测者先前的游戏经验等个体要素都有可能影响测评结果。

总而言之，基于游戏的测评方法存在部分不足之处，但毋庸置疑的是，基于游戏的测评更具有独特的优势，使用游戏作为评估工具是一种日渐重要的方法并具有越来越高的价值。

3 测评范式

3.1 证据中心设计

建立科学有效的测评工具是测量个体心理特性的前提和基础，因此在有关基于游戏的测评的研究中，测评工具的建立和检验是学者关注的焦点之一，证据中心设计为其提供了理论基础，并进一步形成了建立测评工具的范式。

Mislevy 等人最先在 2003 年针对教育评估领域提出概念评估框架(Conceptual assessment framework)——一个用于建立评估的通用模型，由学生模型、证据模型、任务模型、组合模型以及呈现模型五个部分组成(Mislevy, Almond, & Lukas, 2003; Mislevy, Steinberg, & Almond, 2003)，并且包含框架实施的四个过程，分别为呈现过程、响应过程、计分过程和任务选择过程。概念评估框架和四个过程被统称为证据中心设计(Evidence-centered design，ECD)，这是一个更广泛的测量模型，以支持现代化教育评估。证据中心设计同样适用于开发游戏测评工具，Shute 在 2011 年将其概括为三个最核心的组成成分，分别为能力模型、

任务模型和证据模型。

3.2 测评工具建立

第一，定义测量的特质结构，即建立能力模型。能力模型的建构需要研究者根据研究问题确定目标特质，也就是期望测量的知识、技能或者能力、态度，并根据已有理论框架定义目标特质的属性及特征。此模型可以是简单模型，通过任务的完成情况考察某一特质，也可以是复杂模型，在一个游戏中综合考察个体的几种特质(孙海洋, 2011)。

第二，确定反映目标特质的指标及计分规则，即建立证据模型。证据模型是能力模型和任务模型的桥梁，将可观察值汇总并建立预测模型从而推断目标特质。这也是证据中心设计框架最核心的组成成分，可分为统计规则和统计模型两个部分(Shute, 2011)。统计规则的构建在于选择游戏中与能力模型相联系的指标，并设定受测者游戏表现的得分或者得分比率等评分规则，以此得到可观察且可量化的结果。由于游戏的多样性，不同的游戏有不同的数据指标，即使是相同的游戏根据不同的能力模型也可能存在不同的数据指标(温迎, 付玉, 黄贝萱, 2019)。通常来说，指标的选择与确立主要依赖于相关领域的研究基础及研究者的经验与专业知识而无统一标准，其中任务完成时间、关卡完成数量、正确率等是较为常见的数据指标，此外，Nebel 和 Ninaus(2019)提出，借助生理数据可以对玩家的情绪和认知状态有更深入的了解，因此未来研究中可以考虑同时采集相关生理数据作为测评指标。

统计模型的构建在于定义可观察的指标和能力模型之间的关系(冯翠典, 2012)，这种关系可能是逻辑性的，也可能是概率性的。一方面，研究者可以基于简单的计算规则将所选指标的结果得分汇总，直接代表目标特质的水平；另一方面，研究者可以借助贝叶斯网络、随机森林等算法，通过所选指标的结果构建数学模型预测目标特质的水平。统计模型的选择与目标特质、游戏任务、指标数量等因素均有关联。

一般而言，逻辑性的模型较为简单，Vendlinksi 和 Stevens(2002)年设计了一款游戏来评估高一学生的化学知识水平，要求受测者在 23 个不同的情境中识别出指定化学品，受测者可以在游戏中通过实验、查阅书籍等多种方式辅助判断，每种情境计 1 分，得分越高说明化学知识水平越高。DeRosier 等人(2012)对儿童在虚拟社交情境中做出的行为选择赋分并计算总分，评估儿童的社会情绪能力。

而概率性的模型更加复杂，Shute 等人在借助植物大战僵尸(Use Your Brain)游戏预测个体问题解决能力(Problem Solving Skills)的研究中，按照某一种行为占此类行为的比例区间划分等级作为一项数据指标，并建立等级与所测特质水平的概率关系，比如受测者在某一项数据指标上的表现是好时，其在目标特质上表现水平是好的概率为 0.5。结合众多数据指标的表现可依据概率公式预测个体问题解决能力水平。目前尚未有研究结论表明统计模型的类型对测评结果存在显著影响，研究者可以根据研究目的选择合适的统计模型。

第三，设计任务或情境从中获得指标，即建立任务模型。在基于游戏的测评中，游戏即是评估的任务，主要目的在于引出受测者能力的证据，需要定义呈现方式、游戏任务特征、游戏任务的难度和数量、完成游戏任务的可行策略和测试行为的目标水平等(Rupp et al., 2010)。有研究者指出不同类型的游戏涉及不同的技能，可以将电子游戏分成策略类、冒险类、角色扮演类、动作类、模拟类和其他(Dickey, 2006)。其中策略类游戏涉及认知能力、决策能力和战略思维，模拟类游戏与问题解决能力、自我意识以及观点采择相关，角色扮演类游戏则需要想象、合作与计划等特质的参与(DeRosier & Thomas, 2018b)，选择游戏作为测量工具时首先需要考虑目标特质与游戏功能的匹配程度。在研究过程中，研究者可以基于现有游戏提炼有预测作用的指标进行评估，也可以根据研究目的设计新游戏。

利用证据中心设计建立目标特质的测评工具为进一步进行数据采集、处理

与信效度检验提供了必要条件(见图 1)。

3.3 信效度检验

基于游戏的测评作为一种较为新颖的测量技术，信效度的检验更为重要，但是目前相关研究数量少。在当前研究中，游戏测评工具的检验方法与传统心理测验的检验方法类似，具体分为信度检验和效度检验。

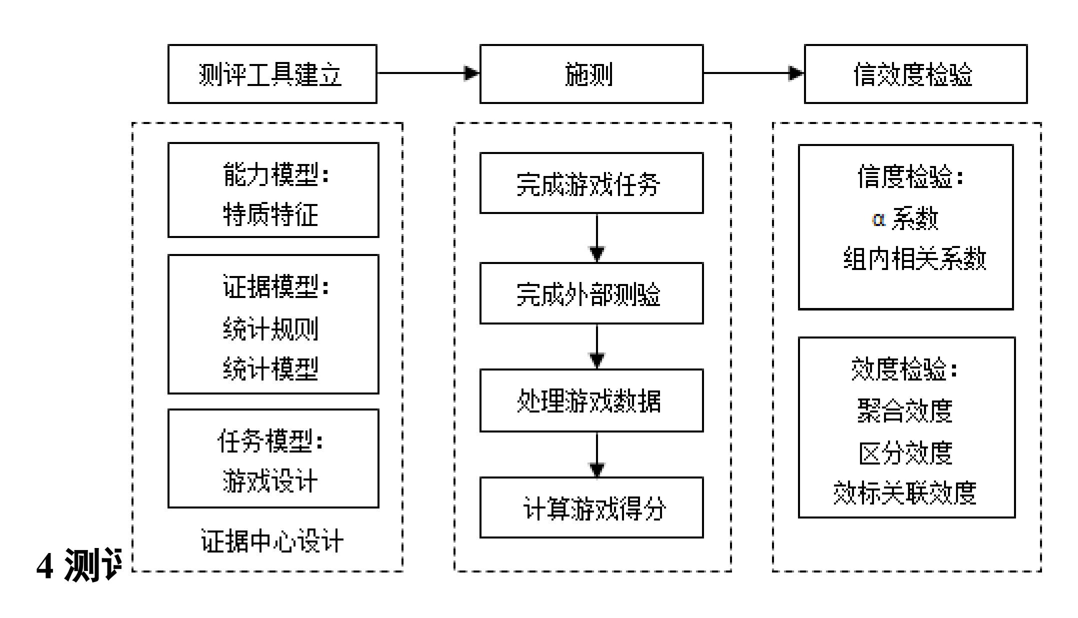
信度，即可靠性(Reliability)，信度系数高即表示该测评工具的结果更一致、稳定与可靠。通常通过计算克隆巴赫(Cronbach's alpha)系数和组内相关系数(Intraclass correlation coefficient)对测评工具的内部一致性信度进行检验(Shute & Moore, 2017)。

以 Kim 等人在 2016 年设计的“物理游乐场(Physics playground)”为例，该游戏共使用 74 个关卡来评估玩家对牛顿三定律的理解，通过控制屏幕上的工具，比如杠杆、斜面等使小球移动到目标位置，若玩家出现使用斜坡、杠杆、钟摆或跳板完成移动目标的行为，则视为其表现优秀。研究者计算了表现优秀的数据的内部相关性(r=0.85)并选择了完成度较高的 29 个关卡进行 α 系数的检验(α=0.87)。此外，研究者还对表现优秀数据的四个结构维度进行验证性因子分析，得到单个维度测量误差小，内部一致性信度高。这些结果在一定程度上说明物理游乐场游戏的信度理想。

效度，即有效性(Validity)，效度高即表示该测评工具能更准确的测出其所要测量的特质。通常借助外部测量工具对聚合效度和区分效度两者进行检验(Rupp et al., 2010)。此外，效标关联效度也是研究者会关注的对象之一，它反映了游戏预测个体在某种情境下行为表现的有效性程度。

以 Weiner 在 2019 年设计的 VR 游戏为例，受测者使用一个头戴控制器和两个手持控制器来完成 3 款 VR 游戏以测量个体的认知能力，包括视觉速度与准

确性、空间想象和视觉追踪能力。测试结束后，受测者需要完成职业能力倾向测验(Employee aptitude survey, EAS)中测量这三种能力的分测验和大五人格测验，此外研究者还获取了受测者的学业成绩(GPA)。将 VR 测试得分分别与外部测验得分和学业成绩建立相关关系和回归方程，结果表明这些测验结果之间存在两两相关且 VR 测试得分可以为学业成绩提供有意义的预测。聚合效度、区分效度及效标关联效度的结果在一定程度上说明 VR 游戏的效度良好。



4 测

在测评范式的指导下，研究者建立了多种游戏测评工具对个体能力与行为开展评估实践。认知能力和非认知能力作为能力的一体两面，对个人发展至关重要(李丽, 赵文龙, 2017)，因此常常成为心理测验测评的对象。目前，基于游戏的测评被广泛应用于个体认知能力的评估(De Klerk et al., 2015)，并且在预测非认知能力方面也有独特的优势。

4.1 认知能力的测评

认知能力是个体在重构和应用知识时所需要的能力，涉及知觉、记忆、注意等基本认知能力和推理判断、想象、问题解决等高级认知能力。基于游戏的测评方法为认知能力的评估提供了新思路，在认知能力评价和认知能力诊断方面均有一定应用。

孙鑫等人(2018)通过推箱子游戏预测个体的推理能力(Reasoning ability)和数学成绩，提取第一步用时占比、完成箱子的比例、思考步数占比、重复步数占比、与最优路径相差步数等 23 个特征建立随机森林模型，并通过计算精确率、查准率和查全率等指标验证模型的预测效果。Shute 等人(2016)借助植物大战僵尸(Use Your Brain)游戏预测个体问题解决能力(Problem solving skills)，包括分析条件和限制、制定解决办法、有效利用资源和工具、监控和调整进程四个维度。根据受测者抵挡僵尸的操作，例如“在有超过五个僵尸时使用能量豆”被认为是有效利用资源和工具的表现，将目标行为/总行为转化为频率后共提取 32 个特征建立贝叶斯网络模型，预测模型得到的结果与瑞文推理测验和模拟投篮任务的得分均存在显著相关。个体的论证推理能力(Argumentative reasoning)也是一种重要的认知能力，研究者使用海上学期 (Seaball—Semester at sea) 游戏要求儿童回答出现的食物是否属于垃圾食品等问题并在多个选项中选择理由，最后对 48 个题目的正确选项进行计分得到游戏总分(Song & Sparks, 2019)，游戏得分越高说明个体的论证推理能力越强。学生在游戏评估中的得分与 CBAL 认知学习能力测验(Cognitively based assessment of, for, and as learning)得分呈中等程度相关，说明了游戏的区分效度和聚合效度；与教师报告的学生成绩与其议论文写作能力上的评级结果也呈显著相关，说明了游戏的效标关联效度。

除了对一般人群的认知能力进行评估外，基于游戏的测评也被用于对认知障碍人群的认知诊断。Manera 等人(2015)采用“厨房与烹饪”的游戏任务评估患有轻度认知障碍和阿尔茨海默病的老人，要求受测者点击屏幕制作菜肴。该游戏分为区分原材料、计划制作工序、实际操作三个过程，涉及感知能力、计划能力和实践能力，最后将完成时间与表现错误次数作为判断指标，受测者游

戏表现与整体认知功能、注意力与思维、执行功能和记忆能力测验的结果均呈显著相关，验证了游戏效度。Flynn 等人(2019)对一个认知障碍夏令营的孩子进行施测和监控，由一组游戏任务组成可以重复测评的认知检测工具，分为感知区分任务(在屏幕上点击正确的目标)和导航任务(通过倾斜 iPad 来引导模拟角色绕过障碍物)，借助自适应算法工具自动记录完成单任务和多任务时的个体情况，共收集 20 项反应指标。通过个体随时间推移的数据结果，可以在改善认知神经障碍的治疗过程中进行更加全面和准确的评估。

4.2 非认知能力的测评

由于反映社会特征及人格特质的非认知能力较难测量，对其关注的时间相对滞后，但随着非认知能力在个体发展中的重要性逐渐展现，近年来，基于游戏的测评在非认知能力测评中的作用也受到了关注。

研究者使用“参观动物园(U ZOO)”这一游戏评估儿童的社会情绪能力，游戏过程中儿童在类似于学校的故事世界中与虚拟角色互动来完成 6 个虚拟社交场景中呈现的情境选择问题，以此评估个体在交流、合作、同理心、情绪调节、冲动控制和社会活动 6 个方面的能力，结果表明 6 个维度的内部一致性均呈正相关。此外研究者获得了受测学生由教师报告的社交技能和行为量表得分、纪律处分和学业适应情况，游戏得分更低的儿童表现出更多社交、行为和学业上的问题(DeRosier et al., 2012; DeRosier & Thomas, 2018a)。此外，对个体团队合作能力(Guenaga et al., 2015)和个人合作行为(Keil et al., 2017)的评估也可以借助游戏实现。

人格特质的测量也受到了学者的重视。Nimwegen 等人(2011)与一家游戏工作室和一家人力资源咨询公司合作开发了一款游戏用于测评个体的依从性，受测者模拟自己在一个公司环境中对发生的事情做决定和表达意见，故事中受测者选择的行动实际上代表李克特四

点量表的分数。Poptropica 岛屿任务游戏可以预测个体的坚持性特质(DiCerbo, 2014)，研究者选择一次通过率低于百分之十的岛屿关卡作为困难关卡，在任务事件上花费的时间以及完成任务的次数作为坚持性的评估指标，并将游戏中三个岛屿任务的两项评估指标建立验证性因素分析模型进行检验，得到各项拟合指数良好。目前心理学领域中较成熟的实验范式——最后通牒游戏(The ultimatum game)可以用于评估个体利他性特质，独裁者博弈游戏(Dictator game)则可以测量个体的公平性特质(Baumert et al., 2014)，通过游戏中行为的表现情况，还可以评估个体主动性攻击和反应性攻击特质(McCreery et al., 2019)。

但在人格特质领域，也有学者得到了不同的结论。Dalveren 等人(2015)借助外科病房导航游戏测试受测者的人格特质，在游戏中受测者需要根据地图达到 10 个不同的目标地点。参照荣格的人格类型(Myers-Briggs type indicator, MBTI)，选择反应时间、行走的距离、走到错误道路的次数、撞墙的次数以及任务成功率等作为游戏行为指标，但是研究者分析了这些人格类型与玩家在游戏过程中的个人表现的相关关系，发现游戏参数与玩家的性格类型之间没有显著的相关性，这对基于游戏的测评是否能有效预测人格特质提出了挑战。

5 未来研究展望

综上所述，本文对基于游戏的测评的概念、范式和实践进展进行了梳理整合，不仅对理解基于游戏的测评这一新兴技术具有重要价值，而且对后续研究的开展也具有重要指导意义。但目前基于游戏的测评的研究仍处于初始阶段，未来学者可以从测评的任务设计、测评的数据分析和测评的实践应用三个角度出发，进一步丰富基于游戏的测评的相关研究。

5.1 测评的任务设计

早期研究多利用一些现有商业游戏如植物大战僵尸、推箱子等（Shute et al., 2016; 孙鑫等, 2018），探讨这些游戏上的表现与某种心理特质的关联。目前有越来越多的研究者尝试根据证据中心设计框架，将游戏机制、游戏内容和内容评估相结合，开发特定主题的游戏（如 Song et al., 2020）。这样在测评相

应的心理特质上更具有针对性，内容更正确完整。一些通用游戏框架的出现，如 Minecraft，也在一定程度上降低了游戏开发的门槛。

但目前大多基于游戏的测评仍采用线性设计模式，针对不同的测试者呈现的游戏情境和内容均一致，这会导致测评需要花费许多时间且评估内容较为单一，因此有研究者提出非线性的游戏模式。一方面表现为分支设计，不同的行为将带来不同的游戏情境。Bacos 等人(2018)在研究中采用一款具有分支故事情节的互动叙事游戏对个体的反事实思维进行测量，由此可见，分支设计也许可以为更高级的特质评估提供方向。另一方面表现为自适应设计，根据对测试者能力的估计，从游戏关卡中选择相应难度水平的游戏，可以大大提高评估效率。Wilson 等人(2006)开发了一款数字竞赛(The Number Race)的自适应游戏软件用于纠正儿童计算障碍，通过评估儿童的计算能力基线并提出适合儿童表现水平的问题，实现对儿童计算能力的训练。尽管此游戏的主要目的在于训练而非评估，但具有一定启示意义。由于目前基于游戏的测评领域中少有研究者进行自适应游戏设计的研究，未来研究者可以参考基于游戏的训练及基于游戏的学习等领域的相关研究进一步探索。

此外，多玩家的大型游戏设计也为同时测量多人及多特质的实现提供了方向。Annetta 等人(2010)开发了一款多玩家教育游戏(Multiplayer educational gaming application, MEGA)以评估 21 世纪数字时代学生的读写能力、创造性思维、执行力和沟通技巧水平，通过观察学生与老师的互动情况、与同伴的讨论情况，以及在玩游戏时的参与程度和花费时间这四类因素进行评价。早期研究对游戏行为的评价及游戏情境的设计均比较简单，计算机技术的发展将为非线性和多人多特质的游戏模式设计带来更多可能性。

5.2 测评的数据分析

早期研究多使用结果数据，目前对过程数据的关注也逐渐增多，倾向于过程数据和结果数据的整合应用。De Klerk 等人在 2015 年总结了 31 项研究成果，其中有 10 个研究使用了游戏结果数据，6 个研究使用了过程数据，其余的两者皆有，这说明研究者对于过程数据的利用率仍有限。随着计算机技术的发展，机器学习在数据处理方面的巨大优势逐渐显现，尤其是通过游戏的方式会得到

数量庞大的数据，传统统计方法无法最大限度提取数据中的信息(Csapó et al., 2012)，而机器学习算法则可以帮助研究者在结果评估阶段建立更复杂的模型。

已有不少研究者引入贝叶斯网络、决策树、随机森林等算法建立预测模型。孙鑫等人(2018)选择测验得分前 25%和后 25%的受测者样本进行特质提取与模型建立，从推箱子游戏中提取 23 个特征指标作为分类数据集的特征值，随机划分训练集和测试集后对数据集进行训练和分类，建立推箱子的游戏表现与推理能力和数学成绩的关系。未来研究也可以考虑结合卷积神经网络处理图像数据，让游戏数据提供更多的信息，以及考虑采用机器学习的非监督学习类型，探究数据内在分组类型或数据各部分的规则，丰富测评结果的分析方法。需要注意的是，尽管这一处理方式具有良好的统计学意义，但是机器学习是数据驱动的建模过程，目的是最大化预测准确性，有时无法兼顾模型中特征本身的意义和结构(Mayer et al., 2014)，仅从数据驱动得到的结论很有可能是没有实际意义的(吴忭, 胡艺龄, 赵玥颖, 2015)，如何在理论基础上与机器学习方法相结合需要更为深入的分析和研究。

5.3 测评的实践应用

在测评内容上，早期基于游戏的测评主要应用于评估个体对知识和技能的掌握程度。不同于采用试卷测试的方法，研究者将考察点融入游戏，使受测者在游戏过程中展现其对知识和技能的理解和应用能力。尤其是评估数学、物理、医疗急救和建筑设计等此类更需要理解应用的知识技能时(De Klerk et al., 2015)，基于游戏的测评是一种有效的工具选择。随着基于游戏的测评的发展，其在认知能力与非认知能力的评估研究中的作用也受到了广泛关注。基于游戏的测评可以在一定程度上避免传统心理测评中存在的易受社会赞许性影响、无过程数据等缺点而受到研究者青睐，越来越多研究者开始借助游戏对个体的心

理特质进行评估与研究，不少企业也自行设计游戏用于人才招聘，以判断应聘者的能力及个性。但如何选择游戏任务中的数据指标代表个体的人格特征困难较高，因此目前对非认知能力的实践较为简单且数量较少，未来研究者在这一方向上进行深入探索。

在应用场景上，近年来基于游戏的测评在临床评估与治疗领域的研究尝试为这一技术带来了新价值。Hautala 等人(2020)开发了一组在线游戏任务用于评估与筛查低年级学生的阅读障碍，Song 等人(2020)设计了一款叫“CoCon”的手机游戏用于评估儿童青少年群体的认知功能并计划将“CoCon”的使用进一步扩展到筛查具有严重认知控制问题的临床人群。基于游戏的测评因其可以建立自动评分系统、详细记录干预期间个体水平变化过程以及通过自适应算法自行调整任务难度而在后续治疗干预中具有较高应用价值。2020 年 6 月，美国食品药品监督管理局批准了一款名为 EndaevorRx 的游戏作为患有儿童多动症孩子的处方药，也反映出这一领域在实践应用中的巨大潜力，如何将理论与实践相结合使基于游戏的测评发挥更大功能需要研究者不懈努力。

在具体应用中，研究者也越来越关注一些细节问题，比如测评指标选择等。DiCerbo(2014)在通过 Poptropica 岛屿任务游戏预测个体坚持性的研究中，预先选择了 4 个行为作为测评指标，但通过小样本测验得到这 4 个指标的关系不稳定且效度较低，因此最终选取了花费在任务事件上的时间、完成任务事件的次数两个数据结果作为测评指标。过于简单地使用两个指标代表个体坚持性的方式会使这一游戏的评估效度令人怀疑，因此在设计证据模型时，需要预先设定适当数量的行为纳入评估指标，并仔细定义行为，证据和结构之间的联系。有研究者指出行为的指标有时无法在证据模型设计初就确定(DiCerbo, 2017)，可

以采用迭代的方式进行识别，从受测者完成任务的过程中对日志文件中哪些元素可以构成证据的假设进行发展和证实。也有研究者关注先前有关游戏测评的默认假设，如游戏对测评动机和参与度的促进，对考试焦虑和学业成绩的影响等(Verma et al., 2019)。

此外，因受测者个体差异带来的测评结果误差也引起了研究者的关注。与女性相比，男性更频繁且持续的玩各种类型的游戏，更熟悉常见的游戏模式、规则等，这可能会帮助他们在游戏中表现更好。研究者曾测量不同的个体差异，例如性别、愉悦感、游戏效能感以及游戏时间等对测评得分的影响，但未得出统一结论(Sanchez & Langer, 2020)。Kim 和 Shute 在 2015 年的研究中比较了男性与女性、经验丰富和经验较少或无的游戏玩家之间的结果差异，得到具有游戏经验的玩家在部分指标上的优势更大，男性的游戏完成率更高且男性与女性在两个重点指标的结果上存在明显性别差异。这些优势程度可能表现于某几个游戏指标上，可以通过测试进行发现与调整优化。基于游戏的测评工具需最大程度地降低玩家个体差异的影响以准确衡量受测者的能力，必要时可以提供充分的学习机会以缩小游戏背景带来的差异。Oranje 等人(2019)则提醒在基于游戏的测评中还需要关注游戏以及玩家的文化环境。将一项在某一文化中设计和验证的游戏应用于另一文化下群体时，应采取跟传统测评类似的评估方式，确保其具有具有跨文化的测量恒等性。

6 结论

基于游戏的心理测评虽然还处于起步阶段，在国内的相关研究也非常少，但可以预见的是，基于游戏的测评在心理测量领域具有巨大的潜力。以证据中心设计为核心的研究范式为测评工具的建立提供了指导，基于该范式在认知能

力和非认知能力方面的研究实践也验证了基于游戏的测评这一技术的有效性。随着计算机技术与游戏技术的不断进步，未来基于游戏的测评有望在教育评价、心理测量和人力资源管理等多个领域发挥重要作用。

参考文献

冯翠典. (2012). “以证据为中心”的教育评价设计模式简介. *上海教育科研, 8*, 12–16.

李丽, 赵文龙. (2017). 家庭背景、文化资本对认知能力和非认知能力的影响研究. *东岳论丛,38*(04) ,142–150.

孙海洋. (2011). 以证据为中心的设计模式及其对口语考试设计的启示. *中国考试, 7*, 20–26.

孙鑫, 黎坚, 符植煜. (2018). 利用游戏 log-file 预测学生推理能力和数学成绩——机器学习的应用. *心理学报, 50*(7), 761–770.

温迎, 付玉, 黄贝萱. (2019). 电子游戏测评综述. *电脑知识与技术, 15*(8), 184–186.

吴忭, 胡艺龄, 赵玥颖. (2015). 如何使用数据：回归基于理解的深度学习和测评——访国际知名学习科学专家戴维·谢弗. *开放教育研究, 25*(1), 6–14.

吴宇. (2015). 浅析严肃游戏在中国的发展. *科技视界, 10,* 135+227.

杨振芳, 孙贻文. (2015). 游戏化招聘：人才选拔的新途径. *中国人力资源开发, 342*(24), 47–52

Annetta, L., Cheng, M.T., & Holmes, S. (2010). Assessing twenty-first-century skills through a teacher-created video game for high school biology students. *Research in Science & Technological Education, 28*, 101–114.

Attali, Y., & Arieli-Attali, M. (2015). Gamification in assessment: Do points affect test performance? *Computers & Education, 83*, 57–63.

Bacos, C. A., McCreery, M. P., & Laferriere, J. R. (2018). Interactive narratives, counterfactual thinking, and personality in video games. *HCI International 2018 – Posters' Extended Abstracts, 850*, 340–347.

Baron, T. (2017). *An architecture for designing content agnostic game mechanics for educational burst games* (Unpublished doctoral dissertation). Arizona State University.

Baumert, A., Schlösser, T., & Schmitt, M. (2014). Economic games: A performance-based assessment of fairness and altruism. *European Journal of Psychological Assessment, 30*(3), 178–192.

Caballero-Hernández, J. A., Palomo-Duarte, M., & Dodero, J. M. (2017). Skill assessment in learning experiences based on serious games: A systematic mapping study. *Computers & Education, 113*, 42–60.

Csapó B., Ainley J., Bennett R.E., Latour T., & Law N. (2012) Technological issues for computer-based assessmen t. In: P. Griffin, B. McGaw, & E. Care (eds), *Assessment and Teaching of 21st Century Skills*(pp. 143–230). Dordrecht: Springer

Dalveren, G.G.M, Cagiltay, N., & Ozcelik, E. (2015). Personality type indicator models in serious games: A case study in a surgical navigation game. *Information Technology Based Higher Education and Training*, 1–4

De Klerk, S., Veldkamp, B. P., & Eggen, T. J. H. M. (2015). Psychometric analysis of the performance data of simulation-based assessment: A systematic review and a Bayesian network example. *Computers & Education, 85*, 23–34.

DeRosier, M. E., Craig, A. B., & Sanchez, R. P. (2012). Zoo U: A stealth approach to social skills assessment in schools. *Advances in Human-Computer Interaction, 1*, 1–7.

DeRosier, M. E., & Thomas, J. M. (2018a). Establishing the criterion validity of Zoo U’s game-based social emotional skills assessment for school-based outcomes. *Journal of Applied Developmental Psychology, 55*, 52–61.

DeRosier M. E., & Thomas, J. M. (2018b). Video games and their impact on teens’ mental health. In: M. Moreno, & A. Radovic (eds). *Technology and Adolescent Mental Health*(pp. 237–253). Cham, Switzerland: Springer

DeRosier, M. E., & Thomas, J. M. (2019). Hall of heroes: A digital game for social skills training with young adolescents. *International Journal of Computer Games Technology, 6*, 1–12.

DiCerbo, K. E. (2014). Game-based assessment of persistence. *Educational Technology & Society, 17* (1), 17–28.

DiCerbo, K. E. (2017). Building the evidentiary argument in game-based assessment. *Journal of Applied Testing Technology, 18*, 7–18.

Dickey, M. D. (2006). Game design and learning: A conjectural analysis of how massively multiple online role-playing games (MMORPGs) foster intrinsic motivation. *Educational Technology Research and Development, 55*(3), 253–273.

Flynn, R. M., Colón-Acosta, N., Zhou, J., & Bower, J. (2019). A game-based repeated assessment for cognitive monitoring: Initial usability and adherence study in a summer camp setting. *Journal of Autism and Developmental Disorders, 49*(5), 2003–2014.

Gamberini, L., Marchetti, F., Martino, F., & Spagnolli, A. (2009). Designing a serious game for young users: The c ase of happy farm. *Studies in Health Technology and Informatics, 144,* 77-81.

Gee, J. P. (2008). Learning and games. In: K. Salen(ed), *Good Video Games and Good Learning*(pp. 21–40). New York: Peter Lang Publishing

Guenaga, M., Eguíluz, A., & Rayon, A. (2015). A serious game to develop and assess teamwork competency. *International Symposium on Computers in Education, SIIE 2014*, 183–188.

Hautala, J., Heikkilä, R., Nieminen, L., Rantanen, V., Latvala, J. M., & Richardson, U. (2020). Identification of reading difficulties by a digital game-based assessment technology. *Journal of Educational Computing Research, 58*(5), 1003–1028.

Heinzen T.E., Landrum R.E., Gurung R.A.R.,& Dunn D.S. (2015). Game-based assessment: The mash-up we've been waiting for. In: T. Reiners, & L. Wood(eds), *Gamification in Education and Business*(pp. 201–207).

Cham, Switzerland: Springer
Keil, J., Michel, A., Sticca, F., Leipold, K., Klein, A. M., Sierau, S., . . . White, L. O. (2017). The Pizzagame: A virtual public goods game to assess cooperative behavior in children and adolescents. *Behavior Research Methods, 49*(4), 1432–1443.
Kim, Y. J., Almond, R. G., & Shute, V. J. (2016). Applying evidence-centered design for the development of game-based assessments in physics playground. *International Journal of Testing, 16*(2), 142–163.
Kim, Y. J., & Shute, V. J. (2015). The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment. *Computers & Education, 87*, 340–356.
Kim, Y. J.,& Ifenthaler, D. (2019). Game-based assessment: The past ten years and moving forward. In: D. Ifenthaler, & Y. J. Kim(eds), *Game-Based Assessment Revisited. Advances in Game-Based Learning*(pp. 3–11). Cham, Switzerland: Springer
Manera, V., Petit, P. D., Derreumaux, A., Orvieto, I., Romagnoli, M., Lyttle, G., . . . Robert, P. H. (2015). “Kitchen and cooking”, a serious game for mild cognitive impairment and Alzheimer's disease: A pilot study. *Frontiers in Aging Neuroscience,* 7–24.
Mavridis, A., & Tsiatsos, T. (2017). Game-based assessment: Investigating the impact on test anxiety and exam performance. *Journal of Computer Assisted Learning, 33*(2), 137–150.
Mayer I., Dierendonck D.V, Ruijven T.V, & Wenzler I. (2014). Stealth assessment of teams in a digital game environment. In: A. De Gloria(ed), *Lecture Notes in Computer Science*(Vol. 8605, pp. 224–235). Cham, Switzerland: Springer
McCreery, M., Krach, S., Bacos, C., Laferriere, J., & Head, D. (2019). Can video games be used as a stealth assessment of aggression?: A criterion-related validity study. *International Journal of Gaming and Computer-Mediated Simulations, 11*, 40–49.
Mislevy, R., Almond, R., & Lukas, J. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series, 1,* 1–29.
Mislevy, R., Steinberg, L., & Almond, R. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research & Perspective, 1*(1), 3–62.
Mislevy, R. J., Behrens, J., DiCerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence centered design, psychometrics, and data mining. *Journal of Educational Data Mining, 4*, 11–48.
Nebel, S., & Ninaus, M. (2019). New perspectives on game-gased assessment with process data and physiological signals. In D. Ifenthaler. & Y. J. Kim. (Eds.), *Game-Based Assessment Revisited* (pp. 141–161). Cham, Switzerland: Springer
Nicholson S. (2015). A recipe for meaningful gamification. In: T. Reiners, & L. Wood(eds), *Gamification in Education and Business*(pp. 1–20). Cham, Switzerland: Springer
Nimwegen, C., Oostendorp, H., Modderman, J., & Bas, M. (2011). A test case for GameDNA: Conceptualizing a serious game to measure personality traits. *16th International Conference on Computer Games*, 217–222
Oranje, A., Mislevy, B., Bauer, M. I., & Jackson, G. T. (2019). Summative game-based assessment. In: D. Ifenthaler, & Y. J. Kim(eds)*, Game-Based Assessment Revisited* (pp. 37–65). Cham, Switzerland: Springer
Rupp, A., Gushta, M., Mislevy, R., & Shaffer, D. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *Journal of Technology, Learning, and Assessme*nt, *8*(4), 1–48
Sanchez, D. R., & Langer, M. (2020). Video game pursuit (VGPu) scale development: Designing and validating a scale with implications for game-based learning and assessment. *Simulation & Gaming, 51*(1), 55–86.
Schoedel, R., Au, Q., Völkel, S., Lehmann, F., Becker, D., Buehner, M., . . . Stachl, C. (2018). Digital footprints of sensation seeking: A traditional concept in the big data era. *Zeitschrift für Psychologie*, *226*(4), 232–245.
Shute, V. J , & Moore,G.R.(2017). Consistency and validity in game-based stealth assessment. In J. Hong, & W. L. Robert (Eds.), *Technology Enhanced Innovative Assessment: Development, Modeling, and Scoring From an Interdisciplinary Perspective*(pp. 31–51). Charlotte: Information Age
Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias, & J. D. Fletcher(Eds.), *Computer Games and Instruction*(pp. 503–523). Charlotte: Information Age
Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior, 63*, 106–117.
Song, H., Yi. D. J.,& Park, H. J. (2020). Validation of a mobile game-based assessment of cognitive control among children and adolescents. *PLoS ONE, 15*(3), 1–18.
Song, Y., & Sparks, J. R. (2019). Measuring argumentation skills through a game-enhanced scenario-based assessment. *Journal of Educational Computing Research, 56(8*), 1324–1344.
StĂNescu, D. F., IoniȚĂ, C., & IoniȚĂ, A. M. (2020). Game-thinking in personnel recruitment and selection: Advantages and disadvantages. *Postmodern Openings / Deschideri Postmoderne*, *11*(2), 267–276.
Turan, Z., & Meral, E. (2018). Game-based versus to non-game-based: The impact of student response systems on students' achievements, engagements and test anxieties. *Informatics in Education, 17*(1), 105-116.
Vendlinksi, T., & Stevens, R. (2002). Assessing student problem-solving skills with complex computer-based tasks. *The Journal of Technology, Learning and Assessment, 1*(3).
Verma, V., Baron, T., Bansal, A., & Amresh, A. (2019). Emerging practices in game-based assessment. In: D. Ifenthaler, & Y. J. Kim(eds), *Game-Based Assessment Revisited* (pp. 327–346). Cham, Switzerland: Springer
Weiner, E. J. (2019). *Cognitive abilitiy in virtual reality: Validity evidence for VR game-based assessment.*

(Unpublished master's thesis). San Francisco State University.
Wilson, A. J., Dehaene, S., Pinel, P., Revkin, S., Cohen, L., & Cohen, D. (2006). Principles underlying the design of "The Number Race", an adaptive computer game for remediation of dyscalculia. *Behavioral and Brain Functions*, *2*(1), 2–19.

Game-based psychological assessment: Conception, paradigm, and practices

XU Junyi, LI Zhongquan

(Department of Psychology, School of Social and Behavioral Sciences, Nanjing University, Nanjing 210023*, China)*

Abstract: Game-based psychological assessment refers to the evaluation of a person's ability, personality and other psychological characteristics through games or gamified activities. It was primarily for the purpose of evaluating learning effects at early period and then developed into the evaluation of psychological characteristics. As a new technology, game-based asssessment has advantages in terms of form, process and outcome. Currently, a paradigm based on evidence-centered design has been developed to design assessment tools and conduct empirical studies. This kind of paradigm is applied to assess individual cognitive and non-cognitive abilities. Future research may focus on task design, data mining, and application.

Key words: game-based assessment, evidence-centered design, cognitive ability, non-cognitive ability